DOCUMENT RESUME

ED 427 058 TM 029 439

AUTHOR Espelage, Dorothy L.; Quittner, Alexandra L.; Kamps, Jodi
TITLE An Application of Generalizability Theory to the Validation of a Behaviorally Anchored Role-Play Measure.
PUB DATE 1998-04-00
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Behavior Patterns; Eating Disorders; Error of Measurement; Females; *Generalizability Theory; *Patients; Psychometrics; *Reliability; *Role Playing; Scores; *Test Construction; Test Theory; Test Use; Test Validity

ABSTRACT

Generalizability theory (g-theory) was used, as an alternative to classical test theory, to evaluate measurement error in a behaviorally anchored role-play measure, highlighting the usefulness of this theory in instrument development. G-theory partitions an observed score into the universe score and error scores associated with separate sources of error variation, and enables the researcher to make decisions about instrument development and validation not afforded by classical test theory. Forty-four female patients at an eating disorders clinic and a sample of 44 young women without eating disorders were recruited for the study. Participants completed questionnaires about demographics and eating disorders and then responded to vignettes about problem family situations. These responses were rated by four raters as part of the development of the behavioral role-play measure of social competence, and g-theory was used to assess agreement among the four raters. Traditional approaches to assessing reliability were compared to generalizability theory to validate the developed instrument. Although four raters were used, results suggest that a single rater could produce a reliable score. Results also show that using g-theory in addition to classical test theory to assess the psychometric properties of the instrument offers a measurement approach that benefits researchers and clinicians. (Contains 5 tables and 28 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Running Head: GENERALIZABILITY THEORY & ROLE-PLAY MEASURE

ED 427 058

An application of generalizability theory to the validation of a Behaviorally anchored role-play measure

Dorothy L. Espelage, Ph.D.

University of Illinois

Alexandra L. Quittner, Ph.D.

Indiana University

Jodi Kamps, B.A.

University of Kansas

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Espelage Dorothy
2A

Paper presentation at the 1998 American Educational Research Association Annual Meeting, San Diego, CA. Address inquiries to Dorothy L. Espelage, Ph.D., Department of Educational Psychology, University of Illinois, 230 Education Building, 1310 South Sixth St., Champaign, IL 61820-6990; 217-333-9139 & espelage@uiuc.edu.

TM029439

BEST COPY AVAILABLE

Psychological measures are an integral part of clinical assessment and are used for diagnosis, the selection of intervention targets and evaluation of treatment outcome (Goldfried, 1982; Peterson & Sobell, 1994). With the advent of health care reform and the greater role of managed care in psychology, researchers and clinicians are being asked to develop specific, short-term, cost-effective treatment approaches that have demonstrated efficacy for particular clinical populations (Hayes, Nelson, & Jarrett, 1987; Korchin & Schulberg, 1981; Wiggins, 1980). This necessitates the development of measures that yield reliable and valid scores that can assess progress during and following treatment in an efficient and meaningful manner (Howard, Moras, Brill, Martinvoch, & Lutz, 1996). For example, changes on standardized measures, such as the Beck Depression Inventory, following a course of treatment are difficult to interpret and fail to identify specifically how an individual's daily life has changed as a result of treatment. Although relatively easy to administer, measures of this type fail to illuminate the processes underlying treatment (e.g., reductions in negative cognitions lead to use of social support network) and therefore fall short in highlighting the specific areas that have been changed as a result of treatment (Quittner, in press; Sechrest, McKnight, & McKnight, 1996).

In contrast, behavioral assessment approaches yield more functional outcomes by more directly measuring an individual's response to specific life situations and by focusing on what a person does in a specific situation (First, Frances, Widiger, Pincus, & Davis, 1992; Kazdin, 1993). Behaviorally-oriented assessment not only improves our evaluation of treatment programs, but also allows us to tailor interventions to particular problems (Haynes, 1990; Nelson & Hayes, 1981). Several techniques are associated with behavioral assessment, including naturalistic observation, role-playing, and self-report measures developed from real-life situations. These approaches typically focus on a range of potential behavioral components

(what the person thinks, does, and says), rather than attempting to measure underlying constructs, such as internal motivations or consistent personality traits. Because the majority of these methodologies involve the use of multiple raters or judges, interrater reliability is a major component of behavioral assessment. With relatively few exceptions, researchers have relied on traditional approaches to assess interrater reliability, including correlations, kappa coefficients, and percent agreement, and continue to estimate other forms of reliability separately (e.g., test-retest, parallel-form). To date, only a handful of studies in psychology have utilized generalizability theory (e.g., Axelrod, Goldman, & Woodard, 1992) to compute reliability of scores derived from measures, and no investigations have used this technique with measures developed specifically for women with eating disorders. The present investigation is the first empirical demonstration of the clinical utility of generalizability theory in validating a behaviorally-oriented role-play measure of social competence in women with and without eating disorders.

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) (g-theory) extends classical test theory to consider multiple sources of measurement error. G-theory partitions an observed score into the universe score and error scores associated with separate sources of error variation, and enables the researcher to make decisions about instrument development and validation not afforded by classical test theory (Shavelson, Webb, Rowley, 1989; Webb, Rowley, Shavelson, 1988). First, variance components are calculated to determine the amount of error introduced by the items, forms, occasions, raters, and the interaction between these factors. Second, these variance estimates are then used to determine which sources of variation need to sampled more rigorously to improve generalizability of scores. Unlike classical test theory approaches that require separate estimation of different types of reliability (e.g., test-

retest), g-theory considers all possible sources of unreliability simultaneously. Third, this analysis estimates the number of items, raters, or occasions needed to produce a reliable score on a measure, allowing the researcher to create instruments with fewer items without jeopardizing reliability.

Thus, this paper reflects a major contribution to the clinical literature as well as to the assessment of eating disorders. Assessment of eating disorders and other clinical populations continue to rely on trait-based measures, do not typically employ multiple raters, and therefore, do not utilize alternative forms of estimating interrater reliability. However, as studies continue to demonstrate that more "targeted" interventions are more effective than generally applied programs (Parloff, London, & Wolfe, 1986; Weist, Ollendick, & Finney, 1991), an increase in behavioral assessment is likely to occur. Subsequently, the use of multiple raters of behavioral observations is also likely to increase. Thus, scholars in the field of counseling psychology must incorporate alternative strategies of psychometric theory and test construction to minimize measurement error and develop measures that can be administered quickly without sacrificing reliability. The present investigation demonstrates the use of generalizability theory, an alternative to classical test theory, to evaluate measurement error in a behaviorally-anchored role-play measure and highlight this theory's utility in instrument development. G-theory is compared to traditional forms of assessing interrater reliability.

Methods

Participants

Forty-four clients, primarily representing consecutive admissions to an outpatient eating disorders program at a local hospital, were recruited for this study on a voluntary basis over a period of two years. The co-therapists of the outpatient program distributed an information letter

and consent form, and intake questionnaires to all group members during an initial interview. Participants were exclusively female, ranging in age from 17 to 27 years (M = 20.70), with a range of 12 to 17 years of education completed (M = 13.75). Forty-three were Caucasian, and one participant was African-American. Forty-two were single, one participant was married, and one was separated. Thirty-four were currently enrolled at a local university, and ten were not students at the time of pre-testing. Twenty-nine participants lived in a dorm or apartment, nine lived with their parents, five lived alone, and one lived with a spouse. Thirty-two participants met the criteria for bulimia nervosa and twelve met the criteria for anorexia nervosa from the Diagnostic Statistical Manual of Mental Disorders (DSM-III-R; American Psychiatric Association, 1987) based on a structured clinical interviews with the co-therapists of the outpatient program.

A sample of 44 non-eating disordered subjects were obtained in the following manner. First, in order to match the eating disordered sample, only female students between the ages of 17 and 27 years were recruited. Female students in four large introductory educational psychology courses were asked to participate in a study investigating the weight and diet concerns of college students. Students who volunteered were asked to complete the Eating Attitudes Test-26 developed by Garner, Olmsted, Bohr, and Garfinkel (1982). The EAT-26 was used as a screening measure, with a score of greater than 20 indicating an eating disturbance or a possible eating disorder (Button & Whitehouse, 1981; Garner et al. , 1982). As an extra check, only students who fell at or below a score of 15 on the EAT-26 were eligible for inclusion in the control sample.

Measures

Participants completed a demographic questionnaire and the Anorexia Bulimia Problem Inventory (Eason, 1983) as part of a longitudinal treatment outcome study. Please see Espelage, Quittner, Sherman, & Thompson (under review) for a detailed description of this study.

Demographic Questionnaire. This questionnaire included questions pertaining to participants' age, years of education, race, living arrangement, and marital status.

The Anorexia and Bulimia Problem Inventory (ABPI). The ABPI (Eason, 1983) is a 42-item behavioral role-play inventory of common problematic situations encountered by women with eating disorders and was developed in a 1983 doctoral dissertation by Eason. In an application of the behavior-analytic approach (Goldfried & D'Zurilla, 1969), Eason conducted four phases of instrument development followed by a validation study. In the Situational Analysis Phase, three women recovering from bulimia, nine with sub-clinical eating disorders, six parents (5 mothers, 1 father) attending a support group, and therapists who specialize in eating disorders provided descriptions of common problematic situations they had experienced or observed in their clients or family members. All were interviewed in person for one hour to identify the specific problematic situations they had encountered in the areas of family life, social life, academics, and eating. These situations were condensed, and 50 items were written as role-play vignettes, including contextual and situational details (i.e., setting, time of day, persons involved). Vignettes were then audiotaped, with a 20-second pause separating each vignette, to allow time for the subject to produce a separate oral response for each item (i.e., What would you say or do in this situation?). In the Response Enumeration Phase, 17 women with sub-clinical eating disorders and 15 non-disordered female college students responded to the 50 vignettes and were then asked to rate each problematic situation using a 4-point scale on

three dimensions: frequency, difficulty, and relevance (e.g., importance). All 50 vignettes were rated as occurring "occasionally" to "a lot"; were "moderately" to "extremely difficult"; and were "important" to "very important." Thus, all 50 items were retained for the next phase of the instrument development.

In the Response Evaluation Phase, judges (e.g., college students, therapists) rated the competence of each response on a 4-point scale (1 = incompetent, 2 = somewhat incompetent, 3 = somewhat competent, 4 = competent). Those items that demonstrated high interrater reliability (using Cronbach alphas > .65) and elicited responses that were distributed across the 4-point competency scale were included in the inventory. Forty-two items were retained and judges' ratings were averaged to create item scores. Total scores were calculated by summing ratings across all items.

A preliminary comparison of the ratings of the eating disordered sample and control group revealed that the sub-clinical eating disordered sample produced significantly less competent responses than controls on 21 of the 44 vignettes, but the groups did not differ on overall ABPI performance. In a validation study including 45 women with subclinical eating disorders and 20 controls, six subscales were created based on the content of the items (e.g., eat/weight concern, family issue) with internal consistencies of each scale ranging from .49 to .89. Results of this validation study indicated that the eating disordered sample had a significantly lower (less competent) mean ABPI Total score than the control group. In general, those with disordered eating produced significantly less competent responses than controls on the Eating and Weight, Male Rejection, and Academic Concerns scales.

Although Eason followed the behavior-analytic model very closely and developed a strong measure, she assessed interrater reliability using alpha coefficients only, which is not a

widely acceptable methodology. Thus, it is not unlikely that there may be some vignettes on the ABPI that may be difficult for judges to reliably rate and may therefore be introducing error. In addition, because she did not conduct a generalizability study, it remains to be answered whether the instrument could be rated reliably with fewer raters and/or fewer items.

Procedures

Data collection. Each participant was seated in a small testing room for administration of the questionnaires and the 42-item ABPI. After completing the questionnaires, participants were presented with a series of audiotaped vignettes that included problematic situations concerning family (e.g., "Parents are upset because you are doing poorly in school."), relationship issues (e.g., "Your boyfriend broke up with you."), school (e.g., "You overslept for an important exam."), and eating or weight (e.g., "Skinny friend changes the subject when you ask her how you look."). They were asked to respond to each situation by describing what they would actually say or do in each situation, and their responses were recorded during a 20-second pause separating each vignette.

ABPI Computerized ratings. Once all 88 participants completed the ABPI, their responses were transcribed by vignette. A text file was created to enter the 88 responses for each of the 42 vignettes, and these responses were then randomized with respect to group membership (ED vs. N-ED). All 3,696 responses were then programmed into a HYPERCARD computer program and rated by four trained raters. Each rater read the vignette and the response, and then independently consulted the ABPI raters' manual to assign a competence rating (1 = incompetent, 2 = somewhat incompetent, 3 = somewhat competent, 4 = competent).

Training of raters. Four undergraduate research assistants rated each of the 3,696 responses. These research assistants were trained by the investigator in a group format to use the

ABPI rater's manual and the HYPERCARD computer program. The experimenter, who was familiar with the ABPI, conducted three 45-minute training sessions. The responses of three eating disordered and three non-eating disordered subjects who were not part of the study sample were used for the training sessions. Raters were trained to a 90% agreement criterion.

To control for rater drift, interrater agreement was calculated using generalizability theory (Cronbach, et al., 1972) after every eight vignettes. Generalizability theory uses analysis of variance (ANOVA) to estimate the magnitude of potential sources of measurement error. Using 42 items and 4 raters to calculate measurement error, g-coefficients ranged from .92 to .95, and phi-coefficients ranged from .90 to .94. In addition, variance due to raters ranged from .09% to .31%. In each g-study, these percentages were small relative to the other variance components, indicating very good agreement among raters.

Interrater Agreement. Three types of analyses were conducted to evaluate interrater reliability: internal consistency coefficients (Cronbach alphas), kappa coefficients, and generalizability studies. First, alpha coefficients were calculated for each vignette using the four raters as items (Table 2). Alpha coefficients ranged from .93 to .98 (M = .94) across the 42 vignettes, indicating good interrater agreement.

Second, Kappa, a conservative measure of interrater agreement that corrects for chance, was the most widely used measure of interrater reliability and was cited in approximately 1,000 articles prior to 1989 (Cohen, 1960, 1968). There are weighted and unweighted kappa coefficients. An unweighted kappa coefficient assumes there are equal intervals between the rating criteria, while a weighted kappa coefficient allows for differences between the criteria. For example, if a pair of raters scored a response as "1" (incompetent) and "2" (somewhat incompetent), and another pair gave the same response a rating of "2" (somewhat incompetent)

and a "3" (competent), the unweighted kappa coefficient would assign the same level of agreement to both situations. In contrast, the weighted kappa coefficient would correct for this situation, by recognizing that a 1-2 discrepancy represents closer agreement than a 2-3 discrepancy. Weightings for each item were created by a psychologist and research assistant who had worked with eating disordered patients, and conducted research in this area.

Although both weighted and unweighted kappa coefficients were calculated in the present study, weighted kappa coefficients were considered more appropriate given the nature of the 4-point rating scale. Because the ABPI criteria was based on an ordinal scale, agreement should be higher when raters use the same end of the scale (i.e., 1-2 incompetent or 3-4 competent), and lower when a 2-3 discrepancy occurs.

More specifically, first raters make a decision about whether a response is competent or incompetent, and then within this general category, decide whether the response is "very"or "somewhat" competent or incompetent. To test this assumption, the frequency of 2-3 discrepancies were calculated over the 42 vignettes. Of the 22,176 non-redundant pairwise rater combinations (528 for each vignette), a 2-3 discrepancy occurred only 957 times (4%) across 42 vignettes. In addition, the number of 2-3 discrepancies were consistent in each vignette. Thus, the 2-3 discrepancy was rarely found and indicated that use of the weighted kappa coefficients was most appropriate.

Weighted kappa coefficients were calculated using the table of weights supplied to the investigator, and as shown in Table 3, these weighted kappas ranged from .66 to .92 (M = .84), and unweighted kappa coefficients ranged from .40 to .83 (M = .72). Given the conservative nature of the weighted kappa coefficient, an a priori decision was made to delete any vignette

with a weighted kappa of less than .60. Accordingly, no vignettes were removed from the measure at this step.

Generalizability-studies. Next, generalizability theory (Cronbach, et al., 1972) (g-theory) was used to assess agreement among the four raters, while considering multiple sources of variation. These analyses provided additional information that was useful in the revision of the measure. If agreement was low then g-theory would provide information about where error is being introduced. G-theory uses analysis of variance (ANOVA) to estimate the magnitude of potential sources of measurement error. In this case, we measured the subjects (object of measurement), and the sources of error (variance) introduced by Raters (r), Items (i), Subject by Item (si), Subject by Rater (sr), Item by Rater (ir), and Subject by Item by Rater interaction or error. These variance components were then used to calculate both a g-coefficient for relative decisions, and a phi-coefficient for absolute decisions. These coefficients are interpreted as reliability coefficients, but have an additional advantage of accounting for multiple sources of error (e.g., items, raters). The relative decisions concerns relative standing of one subject to another, and the absolute decision reflects each subject's score compared to a cutoff score. The total variance in an observed score is equal to the sum of these variance component estimates:

$$v_x = v_s + v_i + v_r + v_{si} + v_{sr} + v_{ir} + v_{sir,e}$$

All 14,784 ABPI ratings were used to conduct the g-study. Variance component estimates, g-coefficients and phi-coefficients are found in Table 4. The percentage of variance due to rater was approximately .28%, which indicated little to no difference between raters' use of the 1 to 4 point scale. In addition, the percentage of variance for the interaction between Subject and Rater was also small (.02%) and suggested raters agreed on the relative standing of subjects. Similarly, the percent of variance due to the Vignette by Rater interaction was small

(.34%), indicating raters used the same part of the scale for each vignette, and introduced little error into the observed score when compared to other variance component estimates. This indicated that raters were in high agreement.

As expected, the largest variance component was due to the Subject by Vignette interaction (66%). Thus, the relative standing of subjects differed from item to item, and those subjects who demonstrated highly competent responses on one item were not the same subjects who gave highly competent responses on other items. Finally, the variance component due to the Subject by Vignette by Rater, error was very small (9%), which reflected the amount of variance explained by other sources of variation and by the Subject by Vignette by Rater interaction.

In the next step, these variance estimates were used to first calculate a g-coefficient and a phi-coefficient using one rater and one vignette. The g-coefficient and phi-coefficient were .18 and .17 respectively. These low coefficients suggested that the score of one rater on one vignette would not be a generalizable estimate of social competence. In order to decrease the amount of error associated with relative and absolute decisions, vignettes and raters should be increased. Thus, a series of g-coefficients and phi-coefficients were calculated manipulating the number of raters and vignettes in the calculations. As can be seen in Table 4, a g-coefficient of .87 and a phi-coefficient of .85 was found with 1 rater and 30 vignettes. Increasing the number of raters and vignettes boosted the reliability only slightly, suggesting that a shorter version of the 42-item measure would yield sufficient interrater agreement, and that one rater could reliably rate the ABPI. For clinical practice, these results indicated that a shorter version of the ABPI could be used by a single clinician, thus decreasing the amount of time required for administration and scoring.

After preliminary item and content analyses, 34 of the original 42 vignettes were retained and clustered on the following four scales: Eating and Weight, Romantic Relationships, Academic Issues, and Family Concerns. Internal consistencies ranged from .67 to .89, representing a significant improvement over those reported in the original Eason study (.49 to .77). Thus, the 34-item version of the ABPI consisted of four internally consistent scales (ABPI-R; Espelage & Quittner, 1996). A final g-study conducted on the final set (34 vignettes) of the ABPI are shown in Table 5.

G-coefficients and phi-coefficients were calculated on the final set of 34 items. As expected, the largest variance was due to the Subject by Vignette interaction (64%), followed by the Subject or object of measurement (20%). The lowest variance estimates were found for the Rater facet, Subject by Rater, and Vignette by Rater interactions. One rater and 34 items yielded a g-coefficient of .90, and a phi-coefficient of .89. These generalizability coefficients increased minimally with additional raters (see Table 5). Thus, the final 34-item version of the ABPI can reliably be scored by one trained clinician.

Discussion

Traditional approaches to assessing reliability were compared to generalizability theory to validate a behavioral role-play measure of social competence raters developed specifically for women with eating disorders by investigating interrater agreement among four trained undergraduate. Cronbach's alpha and kappa coefficients suggested that the four raters demonstrated moderate to high agreement. Although these analyses suggested that reliability was moderate on some vignettes, these findings do not indicate what may be contributing to this lack of reliability. More specifically, the question remained whether the raters introduced this error or whether particular vignettes were more difficult to rate, suggesting that the rating criteria

was confusing. As such, a generalizability study was conducted on the 42 items of the Anorexia Bulimia Problem Inventory and results indicated that the four raters were indeed in high agreement. These findings were replicated in a g-study of the revised ABPI consisting of 34 items.

In addition, the generalizability studies indicated that a shorter version of the ABPI could be used by a single clinician, thus decreasing the amount of time required for administration and scoring, and would yield reliable scores. Although four undergraduate raters were trained to use the ABPI manual and scoring criteria, results of the generalizability study suggested one trained rater could produce a reliably score. Thus, a clinician could be trained to administer the ABPI, use the criteria in the ABPI scoring manual, and score the audiotaped responses in a short period of time. Using g-theory in addition to classical test theory to assess the psychometric properties of the ABPI offers a measurement approach that benefits researchers and clinicians in the pursuit of developing measures to effectively and efficiently evaluate treatment programs.

References

American Psychiatric Association (1987). Diagnostic statistical manual of mental disorders (revised 3rd edition). Washington, DC: Author.

Axelrod, B.N., Goldman, R.S., & Woodard, J.L. (1992). Interrater reliability in scoring the wisconsin card sorting test. The Clinical Neuropsychologist, 6, 143-155.

Button, E.J., & Whitehouse, A. (1981). Subclinical anorexia nervosa. Psychological Medicine, 11, 509-516.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213-220.

Cronbach, L.J., Gleser, G.C., Nanda, A.N., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Eason, B. J. (1983). Social competence and problem-solving abilities in anorexic, bulimic, and non-eating disordered females: Empirical development of a role-playing instrument (Doctoral Dissertation, Indiana University, 1983). Dissertation Abstracts International, 44, 8401599.

Espelage, D., & Quittner, A. (1996). Anorexia Bulimia Problem Inventory – Revised (ABPI-R). Unpublished manuscript.

Espelage, D., Quittner, A., Sherman, R., & Thompson, R. (under review). Integrating assessment and treatment of eating disorders: Validity of a behavioral role-play measure of social competence. Psychological Assessment.

First, M.B., Frances, A., Widiger, T.A., Pincus, H.A. & Davis, W.W. (1992). DSM-IV and behavioral assessment. Behavioral Assessment, 14, 297-306.

Garner, D.M., Olmsted, M.P., Bohr, Y., & Garfinkel, P.E. (1982). The Eating Attitudes Test: psychometric features and clinical correlates. Psychological Medicine, 12, 871-878.

Goldfried, M.R. (1982). Behavioral assessment, an overview. In Bellack, A.S., Hersen, M., & Kazdin, A.E. (Eds.). International handbook of behavior modification and therapy (pp. 81-107). New York: Plenum Press.

Goldfried, M.R., & D'Zurilla, T.J. (1969). A behavior-analytic model for assessing competence. In C.D. Spielberger (Ed.), Current topics in Clinical and community psychology (Vol. 1, pp. 151-196). New York: Academic.

Howard, K.I., Moras, K., Brill, P.L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. American Psychologist, 51, 1059 - 1064.

Haynes, S.N. (1990). Behavioral assessment of adults. In G. Goldstein & M. Hersen, (Eds.), Handbook of psychological assessment (pp. 423-466). New York: Pergamon Press.

Hayes, S.C., Nelson, R.O., & Jarrett, R.B. (1987). The treatment utility of assessment: A functional approach to evaluation assessment quality. American Psychologist, 42,

Kazdin, A.E. (1993). Evaluation in clinical practice: Clinically sensitive and systematic methods of treatment delivery. Behavior Therapy, 24, 11-45.

Korchin, S.J., & Schulberg, D. (1981). The future of clinical assessment. American Psychologist, 36, 1147-1158.

Nelson, R.O., & Hayes, S.C. (1981). Nature of behavioral assessment. In M. Hersen & A.S. Bellack, (Eds.), Behavioral assessment: A practical handbook (pp. 3-38). New York: Pergamon Press.

Parloff, M.B., London, P., & Wolfe, B. (1986). Individual psychotherapy and behavior change. Annual Review of Psychology, 37, 321-349.

Peterson, L., & Sobell, L.C. (1994). Introduction to the state-of-the-art review series: Research contributions to clinical assessment. Behavior Therapy, 25, 523-531.

Quittner, A.L. (in press). Improving assessment in child clinical and pediatric psychology: Establishing links to process and functional outcomes. In D. Drotar (Ed.), Handbook of research methods in pediatric and child clinical psychology. NY: Plenum Publishing.

Sechrest, L., McKnight, P., & McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. American Psychologist, 51, 1065-1071.

Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989). Generalizability theory. American Psychologist, 44 (6), 922-932.

Webb, N.M., Rowly, G.L., & Shavelson, R.J. (1988). Using generalizability theory in counseling and development. Measurement & Evaluation in Counseling & Development, 21, 81-90.

Weist, M.D., Ollendick, T.H., & Finney, J.W. (1991). Toward the empirical validation of treatment targets in children. Clinical Psychology Review, 11, 515-538.

Wiggins, J.S. (1980). Personality and prediction: Principles of personality assessment. Mass.: Addison-Wesley Publishing Co.

Table 1
Demographic Characteristics of the ED and N-ED groups[a]

	ED Group		N-ED Group			
	Mean	SD	Mean	SD	*t* value	χ^2
Age	20.70	2.62	20.25	2.06	.90	
Years of Education	13.75	1.67	14.18	1.26	-1.37	
Race						
% Caucasian	97.73		97.73			
% Black	2.27		2.27			.00
Marital Status						
% Single	95.45		95.45			
% Married/Separated	4.55		4.55			.00
Living Situation						
% Parents	20.45		2.27			
% Dorm/Apt	65.90		81.80			
% Spouse	2.27		4.55			
% Alone	11.38		11.38			8.30

[a] ED = Women with eating disorders group, N-ED = Women without eating disorder group

Table 2

Alpha Coefficients for ABPI vignettes using raters as items

ABPI Vignette #	Standardized Alpha Coefficient	ABPI Vignette #	Standardized Alpha Coefficient
1	.97	**22**	.97
2	.96	**23**	.96
3	.97	**24**	.96
4	.96	**25**	.98
5	.97	**26**	.98
6	.97	**27**	.98
7	.97	**28**	.96
8	.96	**29**	.96
9	.97	**30**	.96
10	.97	**31**	.98
11	.98	**32**	.98
12	.97	**33**	.97
13	.96	**34**	.97
14	.97	**35**	.96
15	.97	**36**	.97
16	.96	**37**	.98
17	.96	**38**	.99
18	.97	**39**	.98
19	.93	**40**	.95
20	.96	**41**	.97
21	.98	**42**	.94

Table 3

Weighted & Unweighted Kappa Coefficients for ABPI Vignettes

ABPI Vignette #	Weighted Kappa Coefficient	Unweighted Kappa Coefficient
1	.83	.67
2	.77	.62
3	.87	.74
4	.79	.63
5	.83	.66
6	.80	.71
7	.85	.73
8	.81	.69
9	.84	.70
10	.81	.68
11	.85	.78
12	.86	.69
13	.82	.72
14	.84	.62
15	.84	.67
16	.79	.50
17	.74	.54
18	.83	.61
19	.66	.40
20	.80	.62
21	.92	.74
22	.76	.60
23	.81	.80
24	.83	.70
25	.85	.81
26	.88	.83
27	.89	.73
28	.83	.68

Table 3 (cont)

Weighted & Unweighted Kappa Coefficients for ABPI Vignettes

ABPI Vignette #	Weighted Kappa Coefficient	Unweighted Kappa Coefficient
29	.83	.69
31	.88	.76
32	.85	.70
33	.84	.73
34	.79	.58
35	.82	.66
36	.82	.68
37	.87	.82
38	.85	.74
39	.86	.76
40	.71	.60
41	.79	.72
42	.73	.58

Table 4

G-Study on 42-item ABPI Data: Variance Estimates, G-coefficients, & Phi-coefficients[a]

	Variance Estimates (% of Total Variance)	G-coeffiecient	Phi-Coefficient
Sources of Variance			
Subject	.18 (16.80%)		
Vignette	.07 (6.90%)		
Rater	.01 (.28%)		
Subject X Vignette	.70 (66.00%)		
Subject X Rater	.01 (.02%)		
Item X Rater	.01 (.02%)		
Subj. X Vig. X Rater, Error	.13 (9.00%)		
Number of Raters & Vignettes			
1 rater, 1 vignette		.18	.17
1 rater, 30 vignettes		.87	.85
1 rater, 40 vignettes		.89	.88
2 raters, 30 vignettes		.88	.86
2 raters, 40 vignettes		.91	.89
3 raters, 30 vignettes		.88	.87
3 raters, 40 vignettes		.91	.90
4 raters, 30 vignettes		.88	.87
4 raters, 40 vignettes		.91	.90
4 raters, 42 vignettes		.91	.90

[a] These calculations are based on all 42 vignettes of ABPI, before deletion of items from scales.

Table 5

G-Study on 34-item ABPI Data: Variance Estimates, G-coefficients, & Phi-coefficients

	Variance Estimates (% of Total Variance)	G-coeffiecient	Phi-Coefficient
Sources of Variance			
Subject	.22 (20.00%)		
Vignette	.07 (6.00%)		
Rater	.01 (.28%)		
Subject X Vignette	.70 (64.00%)		
Subject X Rater	.01 (.02%)		
Item X Rater	.01 (.02%)		
Subject X Vignette X Rater, Error	.11 (10.00%)		
Number of Raters & Vignettes			
1 rater, 1 vignette		.21	.22
1 rater, 34 vignettes		.90	.89
2 raters, 34 vignettes		.91	.90
3 raters, 34 vignettes		.91	.90
4 raters, 34 vignettes		.92	.90